# ATTACHMENT D

Exhibit C

Concept Paper, Thomas Lippert, Norbert Eicker, V0.2, 19.7.2010, confidential

Sketch of a Research and Development Project for the upcoming EC Call 7, "Objective ICT-2011.9.13 Exascale computing, software and simulation"

# Dynamical Exa-Computing (DECO)

Exascale Booster for Cluster Computing


## 1. Background

**Call ICT-2011.9.13**   The FP7 ICT Call 7 will present the "Objective ICT-2011.9.13 Exascale computing, software and simulation". This is the first call in FP7 dedicated to Exascale technology. It will be supported through two or three integrated projects (IP) with a volume of M€ 8 each.

The goal is to develop a small number of supercomputing platforms along with system software, compilers and tools, with the potential of O(100) PF in 2014/15 and clear prospects for Exascale in 2018/19. This activity should go along with the scaling of a few application codes optimized to these platforms reflecting the high potential of computational science and engineering contributing to the solution of today's grand challenges with high social and scientific relevance. Proposals should address extreme parallelism with millions of cores involving algorithms, programming models, compilers, power consumption, etc.

Each integrated project (IP) should comprise one or more SC centres, technology and system suppliers including vendors, industrial or academic centres to co-develop a small number of Exa-scaled application codes. It is expected that about 40 % of the budget will be foreseen for (application) software and 60 % for architectural and system development.

Proposals may include international cooperation components. All software should be developed under the open source paradigm. The IP should involve at least three countries from Europe. EC officers have confirmed that US companies may be involved and that the choice of contributors is determined by the requirements of the project alone and does not need to be politically equilibrated.

The date of publication of the call will be the September 28, 2010, as submission deadline, January 18, 2011 is foreseen so far.

**Many Core CPUs**   The recent announcement of Intel to develop the Many Integrated Core-line of processors (MIC), starting with the MIC *Knights Ferry* processor and presenting the MIC *Knights Corner* processor until 2012, promises to start a new era of co-processing for cluster computers. Intel's significant investment in compiler technology and node-parallel programming paradigms (Ct) will provide the X86 programming model. Thus, a relatively continuous migration towards many core computing is feasible in contrast to GPU based accelerator models.

**Hybrid Computing**   *Fine-grained hybrid computing* or *local hybrid computing* using standard processors joined by accelerators can considerably speed up quite a few applications. However, as suitable applications are characterized by a very low (or better no) node-wise
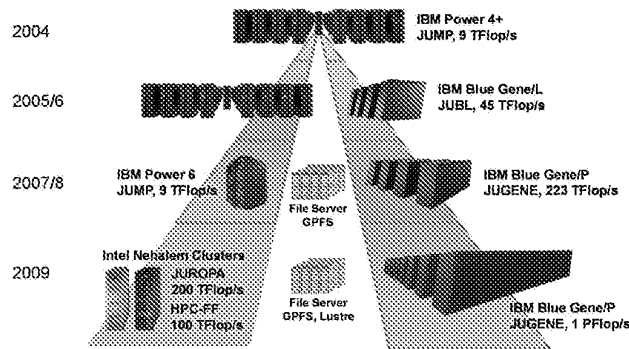
ParTec_00001335

Exhibit C

Concept Paper, Thomas Lippert, Norbert Eicker, V0.2, 19.7.2010, confidential

parallelism at all,[1] these shortcomings, for most of the relevant applications, appear to prevent local hybrid computing from being a scalable architectural model, reaching out to Exaflop/s.

**Code Analysis**   The Jülich Supercomputing Centre (JSC) introduced two lines of architectures in 2005 to better match the characteristics of the application portfolio. On the one hand, the JSC implemented a highly scalable IBM Blue Gene system (JUGENE) for highly scalable codes (Highly scalable codes, sparse-matrix vector like or dominated) and finally in 2009 a flexible general purpose cluster system (JuRoPA) for complex codes with large local memory requests (adaptive grids or coordinate based, all-to-all or more intricate communication patterns, large memory).  So far the distinction between scalable and complex was made on the code level.



A closer analysis of the characteristics of the portfolio of scientific HPC application codes reveals that many codes with future Exascale needs include, on the one hand, code blocks that are well suited for Exascaling, and, on the other hand, such code blocks that are too complex to be so scalable. In the following, the distinction between highly scalable and complex is made on the level of code blocks, and we introduce the notions Exascale Code Blocks (ECB) and Complex Code Blocks (CCB).

Obviously, there is no purist's highly scalable code, but there is no strictly complex code as well. Each code has highly scalable and less scalable complex elements. In fact, there is a continuum between both extremes. Interestingly, many less scalable elements of a code do not require high scalability but instead require large local memory. It is also evident that all-to-all communication elements have a high advantage under smaller parallelism. The question is raised: can we adapt the hardware architecture of future systems to take benefit from this situation?

For such problems where a decent balance between ECBs and CCBs is given in terms of the relative amounts of memory (i.e. the degrees of freedom handled in of ECB vs. the CCB), execution times and data to be exchanged, it suggests itself to adapt to this situation by means of a specific architectural solution, consisting of a traditional cluster computer approach along with an Exascale booster connected through the cluster's network. This dualistic approach has the potential to widen the anticipated narrow application field of pure Exascale systems substantially.

**Exa-Scale Booster**   A *coarse-grained* architectural model emerges, where the highly scalable parts or ECBs of an application code are executed on a parallel many-core architecture, which is

---

[1] For most applications benefiting from hybrid computing, parallelism is restricted to the parallelism intrinsic to the accelerators like GPUs or many core processors.

Exhibit C

Concept Paper, Thomas Lippert, Norbert Eicker, V0.2, 19.7.2010, confidential

accessed dynamically, while the CCBs are executed on a traditional cluster system suitably dimensioned, including the connectivity along with a refined dynamical resource allocation system.
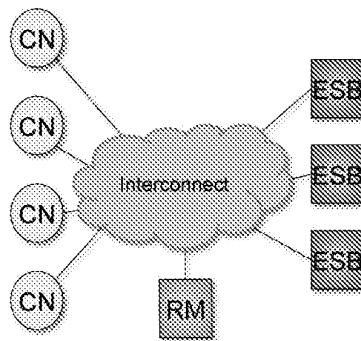
In this conceptual paper, a project for the development of a cluster computer with many-core co-processing elements, called Exascale Booster (ESB), is proposed as a viable model achieving Exascale in 2018.

## 2. Coupled Cluster-Booster Architecture

**Future of Clusters**  Within the period from 2010 to 2014 the speed of commodity processors is expected to increase by about a factor of 4 to at most 8 as the mantissa of Moore's Law appears to decrease in recent years.[2] Can the next generation of cluster computers compete with proprietary solutions like Blue Gene /Q? The next generation of IBM Blue Gene systems, appearing end of 2011 will achieve a factor of 10 higher performance per processor that the Blue Gene /P systems from 2008 at the same power envelope owing to many-core technology. In addition, highly refined compiler technologies will improve the BGQ performance substantially. In order to compete with these developments clusters are forced to utilize accelerator technologies explicitly as accelerators are not expected to be integrated into commodity processors before 2015.

Clusters at Exascale will require virtualization elements in order to guarantee resilience and reliability. While local accelerators, in principle, allow for a simple view on the entire system and in particular can utilize the extremely high local bandwidth, they are absolutely static hardware elements, well suited for farming or master-slave parallelization. Hence, it would be difficult to include them in a virtualization software layer. In addition, there would be no fault tolerance if an accelerator fails, and there was no tolerance for over or under subscription.

**Architecture**  A sketch of a cluster coupled to an ESB is given by the following figure:



The cluster's compute nodes (CN) are internally coupled by a standard cluster interconnect, e.g. Mellanox InfiniBand. This network is extended to include the booster (ESB) as well. In the figure we have drawn three such boosters. The ESBs each consist of a multitude of many-core accelerators connected by a specific fast low-latency network.

---

[2] Source: IPC 2010, Hamburg, talk by Prof. Erich Strohmeier.

Exhibit C

Concept Paper, Thomas Lippert, Norbert Eicker, V0.2, 19.7.2010, confidential

This connection of the CNs with the ESBs should be very flexible. A sharing of accelerator capability between compute nodes becomes possible. The virtualization on the cluster level is not hampered by the model and the full ESB parallelism can be exploited. The ESB-to-CN assignment proceeds via a dynamical resource manager (RM) as developed by the group of Prof. Wolf at GRS. A static assignment at start-time can be made dynamic at run-time. All CN-ESB communication proceeds via the cluster network protocol. The Intra-AC communication will require new solutions. The ESB allocation can follow the application needs and fault tolerance is guaranteed in case of accelerator failures while all compute nodes share the same growth capacity.

**Booster**    As compute element of the booster Intel's many-core processor Knight's Corner (KC), to be available in 2012, is proposed. The KC-chip will consist of more than 50 cores and is expected to provide a DP compute capacity of over 1 Teraflop/s per chip. With 10.000 elements a total performance of 10 Petaflop/s would be in reach in 2012. The predecessor of KC, the Knight's Ferry processor (KF) will be used in the project to create a PCIe-based pilot system for to study the cluster-booster (CN-ESB) concept already in autumn 2010.

As the compute speed of KF exceeds current commodity processors by a factor of about 10, the intra-ESB communication system has to be dimensioned accordingly. Today's QDR IB provides a data rate of about 80 Gigabit/s per card (or node) (duplex). The ESB's communication system requires at least 1 Terabit/s per card (duplex). A promising candidate to such a performance is the communication system EXTOLL as developed by the group of Ulrich Brüning at Mannheim University. EXTOLL provides a communication rate of 1.44 Terabit/s per card. It realizes a 3d topology providing 6 links per card. Concerning its simplicity, this topology appears to be the ideal one for a booster based on many- core accelerators. Even with 2 directions reserved for cut-through routing, EXTOLL can saturate the QPI performance as far as the data rate is concerned. The latency can reach 0.3 μs, when based on an ASIC realization. Currently, EXTOLL is realized by means of FPGAs.

Such concept for the booster requires several novel and innovative developments. Access to the Intel QPI will be essential to achieve a bandwidth per node beyond 1 Terabit/s. Besides the development of specific boards for the KCs, the internal and the external communication system, an enabling middle-ware layer will be an essential component of the booster. On top of this, the cluster-booster system requires the development of an adapted compiler technology, the introduction of specific programming models, and the development of very fast math libraries.

## 3. Project and Consortium

**Consortium**    In compliance with the ideas of the EC as announced on June 11, 2010, for the "Objective ICT-2011.9.13 Exascale computing, software and simulation" the following project consortium might be proposed:

| | | |
|---|---|---|
| **Coordinator:** | **JSC** (D) | Intel-ParTec-FZJ Exacluster Lab in cooperation with the GRS |
| **Further Centres:** | **BSC** (E) | Programming models and compiler technology |
| | **LRZ** (D) | Many-core utilization (tbc) |

Exhibit C

Concept Paper, Thomas Lippert, Norbert Eicker, V0.2, 19.7.2010, confidential

| | | | |
|---|---|---|---|
| **Companies:** | **Intel** (D, F, B) | Knight's Ferry and Knight's Corner design, QPI, X86, simulations |
| | **ParTec** (D) | Cluster operation and communication system, dynamical scheduling |
| | **Mellanox** (IL) | IB, Inter-cluster communication, cluster-to-accelerator communication |
| | **EuroTec** (IT) | Integrator, energy-aware cooling system (tbc) |
| **Universities:** | **Heidelberg** (D) | EXTOLL, Inter-accelerator communication |
| | **Leuven** (B) | Space Weather application code (tbc) |
| | **Lausanne** (SL) | Human Brain Simulation (tbc) |
| | **CERFACS** (F) | Fluid Engineering (tbc) |
| | **Barcelona** (E) | Computational Biology (tbc) |
| | **Regensburg** (D) | Outreach to SFB Hadron Physics with special |
| | **Wuppertal** (D) | purpose Exascale machine development (QCD) |

**Project**

**Phase I (2011-2012)**    Start with a 64 node KF system based on EXTOLL and PCIe communication.

Connect ESB via Mellanox-IB to smaller host cluster.

Start developing dynamical RM and allocation scheme.

Start work on applications using pilot system.

**Phase II (2012-2013)**    Construct prototype based on KC, EXTOLL, IB with O(250) AC-nodes.

Continue work on software and applications.

Build Cluster-Booster system with O(10.000) nodes mid 2013 (external funding).

So far, the final funding scheme is not known. The EC is indicating 8 M€ for each of three projects, where some institutions like the centres usually have to contribute. The possibility was mentioned that the EC will consider only two projects with 12 M€ funds for each project.